

# Dynamic Ising Model: Reconstruction of Evolutionary Trees

P.M.C. de Oliveira

Instituto de Física, Universidade Federal Fluminense  
Av. Litorânea s/n, Boa Viagem, Niterói 24210-340, RJ, Brazil  
and National Institute of Science and Technology for Complex Systems  
e-mail address: pmco@if.uff.br

## Abstract

An evolutionary tree is a cascade of bifurcations starting from a single common root, generating a growing set of daughter species as time goes by. Species here is a general denomination for biological species, spoken languages or any other entity evolving through heredity. From the  $N$  currently alive species within a clade, distances are measured through pairwise comparisons made by geneticists, linguists, etc. The larger is such a distance for a pair of species, the older is their last common ancestor. The aim is to reconstruct the past unknown bifurcations, i.e. the whole clade, from the knowledge of the  $N(N-1)/2$  quoted distances taken for granted. A mechanical method is presented, and its applicability discussed.

PACS: 02.10.Ud; 89.75.Hc; 87.23.Kg.

Natural experiments [1] are those where the experimenter cannot manipulate the object of study. Only comparisons can be made. It is a recent field of research allowing quantitative studies of historical evolutions. Figure 1 exemplifies a clade. At left, the traditional cladogram showing the successive speciations. This kind of draw is familiar to geneticists, linguists, etc. A good historical description entitled *Trees before and after Darwin* was recently published [2]. At right, on Figure 1, we add the not so familiar concept of ultrametric distances on which our analysis is based. A direct measure of such a distance demands scarce fossil data. However, researchers perform indirect measurements of such pairwise distances by comparing features of currently alive species. How to perform these measurements is a vast field of research, out of the current scope. A good review can be found in [3] and references therein. Human evolution can be traced back with genetic or linguistic measurements [4]. A further, still incomplete list of works is shown in [5] for genetics, and [6] for linguistics. Here, we simply suppose the pairwise distances of a given clade were measured among its  $N$  current alive species, i.e. a set of  $N(N-1)/2$  positive numbers taken for granted. The purpose is to reconstruct the whole tree from these data.

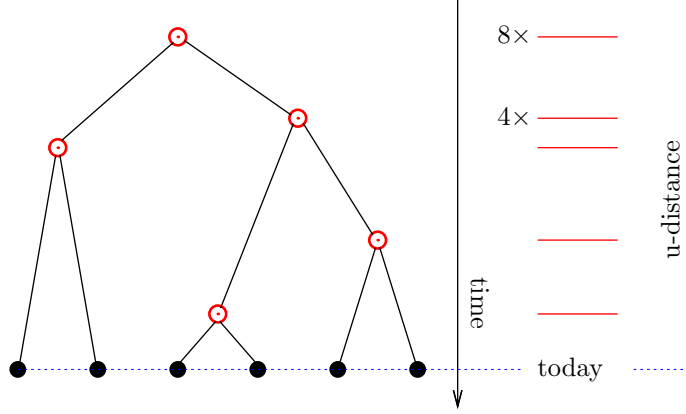


Figure 1: A clade: Closed circles represent the 6 known currently alive species, open circles their past ancestors. The ultrametric distance between two alive species is the time counted from today back to their last common ancestor: 15 distances  $D_{ij}$  are shown in the spectrum at right. The uppermost level is 8-fold degenerate, i.e. the same distance appears 8 times. It corresponds to the root, the single original species. The second level is also 4-fold degenerate. The whole clade (family) is divided in two sub-clades (genera), the 4 rightmost currently alive species belong to one genus, the 2 leftmost to a second genus.

In short, by knowing the distances exemplified at right in Figure 1, the problem is to draw the corresponding tree at left. A mechanical solution follows, each currently alive species is associated to a unitary mass particle moving along a  $X$  axis. Particles are transparent, they can pass through each other. All  $N$  particles are initially released at the origin  $x = 0$  with random velocities (zero sum, keeping the center of mass at rest). Particles interact through the energy

$$E = \frac{1}{2} \sum_{ij} J_{ij} (x_i - x_j)^2 \quad (1)$$

summed over all pairs  $(i, j = 1, 2, \dots, N)$ , with coupling constants given by

$$J_{ij} = D - D_{ij} \quad (2)$$

where  $D$  is an adjustable parameter (we will get rid of it soon).  $D_{ij}$  are the quoted distances. The movement follows Newton's law, the accelerations

$$\ddot{x}_i = - \sum_j J_{ij} (x_i - x_j) \quad (3)$$

form a set of  $N$  linear, second order differential equations which can be solved by diagonalizing its corresponding  $N \times N$  secular matrix. Before that, let's foresee the movement.

Take  $D$  in between the two uppermost levels, Figure 1. If alive species  $i$  and  $j$  belong to the same genus, the coupling constant  $J_{ij}$  is positive (attraction). Otherwise,  $J_{ij}$  is negative (repulsion). The two genera repel each other, while attraction holds inside each genus. Therefore,  $n$  ( $N - n$ ) particles belonging to one (the other) genus remain clustered running away towards one (the other) sense along the  $X$  axis. The eventual partition defines two genera. The same process is repeated within each just discovered genus, and so on, reconstructing the whole clade.

The secular matrix of Equation (3) can be divided as

$$\mathbf{S} = N D \mathbf{I} + D \mathbf{G} \quad (4)$$

where  $\mathbf{I}$  is the identity and  $\mathbf{G}$  is a  $N \times N$  matrix with all entries  $G_{ij} = 1$ . Matrix

$$\mathbf{S} = \begin{pmatrix} \sum D_{1j} & -D_{12} & -D_{13} & \dots \\ -D_{21} & \sum D_{2j} & -D_{23} & \dots \\ -D_{31} & -D_{32} & \sum D_{3j} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} \quad (5)$$

does not depend on (hereafter discarded)  $D$ , only on the measured distances  $D_{ij}$ . Their  $N$  eigenvectors completely define the movement. Among them, two deserve particular comments, two next paragraphs.

The Goldstone eigenvector has  $N$  unitary entries  $(1, 1, 1 \dots 1, 1)$ , whose eigenvalue is always null. Interaction energy (1) presents only internal forces between the  $N$  particles themselves, thus the global center of mass remains at rest at the origin, i.e.  $x_1 + x_2 + x_3 \dots x_{N-1} + x_N = 0$ . Furthermore, any other eigenvector  $(a_1, a_2, a_3 \dots a_{N-1}, a_N)$ , is orthogonal to this always-present Goldstone, i.e.  $a_1 + a_2 + a_3 \dots a_{N-1} + a_N = 0$ . In other words, any eigenvector besides the Goldstone is a series of positive and negative entries with zero sum. Matrix  $\mathbf{G}$  nullifies all these further eigenvectors.

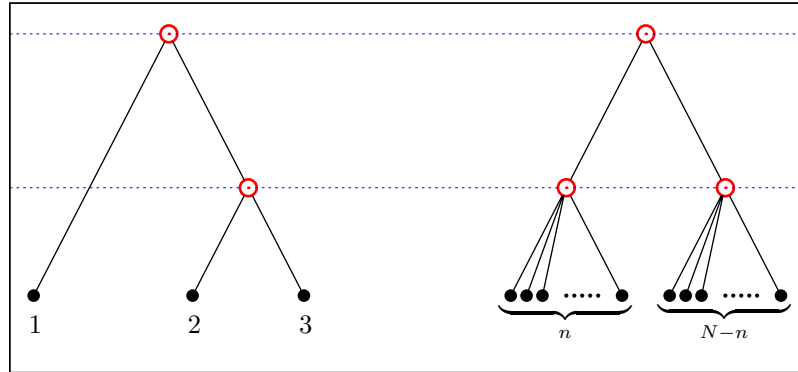


Figure 2: Some analytically solved trees.

Among them, the eigenvector with highest eigenvalue presents  $n$  positive entries for one genus,  $N - n$  negative entries for the other genus, thus solving our problem. Let's call it the partition eigenvector. For the simple clade shown at left in Figure 2, for instance, the partition eigenvector is  $(2, -1, -1)$ . For the largest clade at right, it is  $(N - n, N - n, N - n \dots - n, -n)$ .

The other  $N - 2$  eigenvectors are unimportant, governing only the internal movement within each already-separated genus. In short, given some  $D_{ij}$  dataset, the only task is to compose matrix (5), finding its eigenvector with highest eigenvalue. The signs of its entries describe the correct partition.

However, measured distances suffer from drift fluctuations. Therefore, they do not reflect the exact degeneracies. Each degenerate level becomes a band of neighboring levels, no longer degenerate. In Figure 1, the highest and second highest levels would be represented by two bands with 8 and 4 separated levels, respectively.

Once a given bifurcation was successfully reconstructed, the corresponding broken degeneracy can be restored as follows. Let's  $n_1$  and  $n_2$  be the number of species in each branch of the quoted bifurcation. Then, the  $n_1 n_2$  corresponding distances displayed in the measured dataset represent indeed different measurements of the same single value, namely the real ultrametric distance. It is, then, better represented by the average over measurements: one replaces all  $n_1 n_2$  distances in the  $D_{ij}$  original dataset by their average, restoring the corresponding degeneracy. (Besides the average, dispersion serves to estimate the age uncertainty.)

The band widths increase with the total evolutionary time, due to accumulated random drift. If the clade under study is too old, these bands tend to overlap over each other, and the model may fail beyond some degree of randomness, as any other method. How is it robust against these fluctuations?

Hereafter, we analyze the method performance in these real situations. Our strategy is simple: to test the method with clades for which one knows the entire past history since the first bifurcation. We construct these clades in a computer, following two ingredients [7]. First, one starts from a single species. With a small fixed probability  $b$ , at each new time step the species can bifurcate. After that, each of the two emerging species evolves independent of the other. New branches may also bifurcate. One can book the exact times when each bifurcation occurs. An example of such a tree is shown in Figure 3.

Second ingredient, each species' internal characteristics are represented by a sequence of  $L$  bits 1 or 0. At each time step, this bitstring is mutated, i.e. an average number  $m$  of its bits are randomly chosen and inverted from 0 to 1 or vice versa. When some species bifurcates, its current bitstring is copied to each new branch, both suffering independent random mutations thereafter. The scaling between  $m$  and  $L$  sets the maximum evolutionary time (time-back horizon) one can hope to reconstruct with the available accuracy. We use different seeds for random number generators governing the bifurcations ( $R1$ ) or mutations ( $R2$ ). Keeping the same  $R1$  for different computer runs with different values of  $m$ , one can test the same tree topology under different mutation rates.

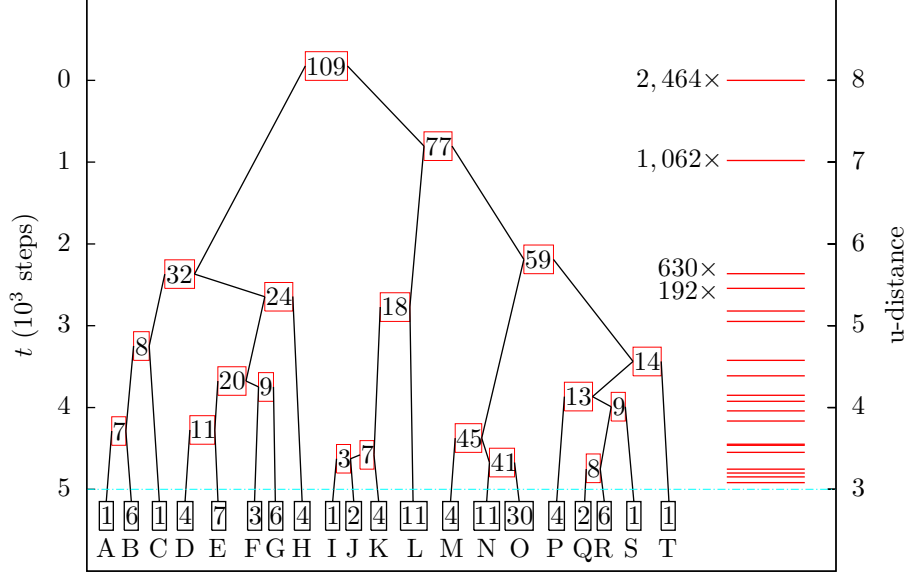


Figure 3: Computer generated evolutionary tree. Starting from a single species, a first bifurcation occurs at  $t = 0$ . A discrete clock  $t = 1, 2, 3 \dots$  runs downwards. Each branch (species) can bifurcate with probability  $b = 0.0005$  at any time. When a bifurcation occurs, the number inside the corresponding square shows the quantity of alive species below it at  $t = 8,000$  (today, not shown) when one counts a total of  $N = 109$  alive species in this particular realization. Bifurcations occurred beyond  $t = 5,000$  are not shown for clarity.

The distance between two bitstrings is the number of unmatched bits divided by  $L/2$  (the random expected value) for normalization. Some similar normalization procedures are followed by linguists in order to cancel out phonetic accidental coincidences [8]. The maximum distance should be 1. However, due to fluctuations, old clades may present some distances slightly larger than 1, say a percentage  $P\%$ . They are of course statistically meaningless data. A nearly equal quantity below 1 is also supposed to be meaningless. Thus,  $2P\%$  is a first, crude estimate for the dataset degree of distrust. Another approach would be the analysis of never-touched-bits, how this set still holding the original information shrinks as evolutionary time goes by [9].

Many clades like that exemplified in Figure 3 were tested, the result is indeed the expected one: perfect reconstruction up to a certain degree of randomness. The general behavior can be appreciated in Figure 4, showing (the top part of) the eigenvalue spectrum of matrix (5) — not to be confounded with the spectrum of distances. At left, the reconstruction is perfect, while the highest eigenvalue is still separated from the band below it, thus corresponding to the correct partition eigenvector. Note also the saturation of the whole spectrum near

the largest conceivable eigenvalue  $N$ , for increasing randomness. The correct partition is obtained up to  $m = 4.1$ , where approximately  $2P = 68\%$  of the whole dataset is statistically meaningless. For  $m = 5.1$  with  $2P = 77\%$ , the partition eigenvalue is surpassed by some other competitor, and the reconstruction fails. As a rule-of-thumb for real  $D_{ij}$  datasets, they can be safely considered trustable if: I) The highest eigenvalue of matrix (5) is separated from the others; and II) At least the second largest eigenvalue is smaller than the upper limit  $N$ .

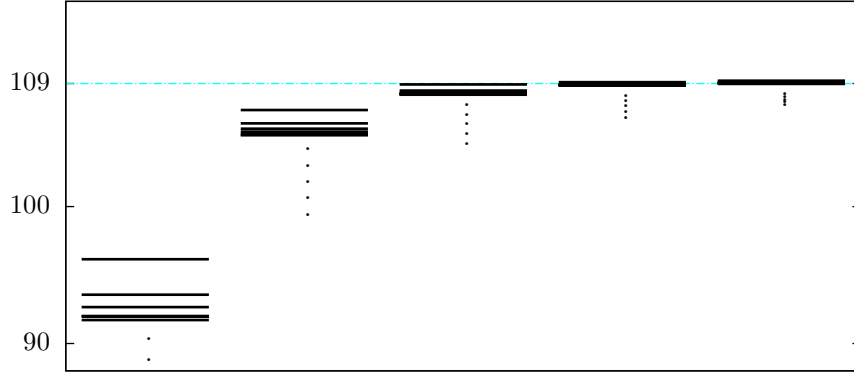


Figure 4: Six topmost secular eigenvalues with increasing fluctuation ( $m = 1.1, 2.1, 3.1, 4.1$  and  $5.1$  from left to right,  $L = 16,384$ ), for the same tree in Figure 3. Other 103 smaller eigenvalues are not shown (indicated by dots). For increasing fluctuation, the whole spectrum shrinks and saturates near the maximum conceivable value  $N = 109$  (except the always-present null Goldstone eigenvalue). Thus, in the limit of large fluctuations, the topmost eigenvalue becomes no longer isolated from the band below it, the genera partition may become wrong. In the current case this occurs at right, for  $m = 5.1$ , when the 2,464 distances corresponding to the highest level in Figure 3 form a wide band with only 16% of the remainder 3,422 data below it, as they all should be.

In hard cases like the rightmost in Figure 4, the remote past of the first bifurcations is inaccessible with the accuracy at disposal. The measured dataset presents too much statistically meaningless entries (distances near 1). They can be gradually expunged as follows. All  $N(N - 1)/2$  links between species are initially present, forming a completely connected network with  $N$  vertices. One cuts the link corresponding to the largest distance  $D_{ij}$ , then the second largest, and so on. At some point along this sequence, the network becomes disconnected in two pieces. The very old history about this separation is inaccessible. But the recent history is not lost. The secular matrix method is then applied to each piece separately (expunged internal links re-included). If the above criteria I and II are not fulfilled for one piece, the cut-link procedure continues within it. Following this strategy for the tree in Figure 3 with  $m = 5.1$ , the isolated

species T is first disconnected from the other 108. Then, also isolated species S disconnects, followed by A and by C. The set of 105 remainder species still does not fulfill the correctness criteria I and II. Next, the block BH with 10 species disconnects, and can be successfully reconstructed by the method. The remainder 95 species do not. Continuing the cuttings, block IJK disconnects, then P, then FQR, then G and L, all successfully reconstructed, and so on.

After each successful step where a correct bifurcation partition is found (under criteria I and II), the corresponding degeneracy is restored, changing the original dataset. Thus, one can re-start the whole process using the new dataset. Degeneracies are, thus, hierarchically and gradually restored. They can also be partially restored, by combining already defined blocks (groups among A, B, C ... R, S, T). After many runs, in case the complete reconstruction still fails, one can observe which blocks are most responsible for failures (normally isolated species or small sub-clades), and remove them from the dataset. Within the hard case  $m = 5.1$  in Figure 3, by removing isolated species A, C, S and T, and also blocks H, L and PQR, the remainder main tree with  $N = 78$  species could be correctly reconstructed. Isolated, each removed block is also correctly reconstructed, but one cannot know where or when it should be branched from the main tree, because this occurred before the time-back horizon at disposal.

There are a lot of alternative methods to reconstruct evolutionary trees [3, 10]. None of them considers the degeneracies appearing in the spectrum of real, ultrametric distances (past time since the last common ancestor), neither the breaking of these same degeneracies due to evolutionary drift. This feature distinguishes the present method, the quoted degeneracies are gradually and hierarchically restored.

Also, the model dynamics considers *all*  $N(N - 1)/2$  distances at each step, mitigating the effects of the statistically meaningless part of the data. This feature is absent from traditional methods as the pioneering UPGMA, where the pair of species corresponding to the smallest distance is joined into a single species, whose distance to each remainder species is the average between both former distances. Instead of simply choosing the smallest distance, minimum-evolution approaches minimize quantities involving all distances, like our method, so improving the performance. Nevertheless, being yet neighbor-joining recipes, the number of distances is always reduced by  $N - 1$  at each step. Reference [10] presents such a method and comparisons with others. The best performances are equivalent to ours. Indeed, in many tests, whenever the correct partition is obtained by the method in [10], it is also obtained by the current one. The tree in Figure 3, for instance, is correctly reconstructed up to  $m = 4.1$  by both methods, but both fail under a little bit larger degree of randomness,  $m = 4.2$ . This coincidence indicates that reconstruction correctness is limited only by fluctuations in the measured dataset, not by drawbacks of the methods themselves. Moreover, when both fail, two different (wrong) partitions were observed. A posteriori, their comparison serves as a further, final criterion for reconstruction correctness.

The author is indebted to Søren Wichmann and José Soares de Andrade Jr. for critical readings of the manuscript and helpful suggestions.

## References

- [1] J. Diamond and J.A. Robinson (eds.), *Natural Experiments of History*, Harvard University Press, Massachusetts (2011).
- [2] P. Tassy, *J. Zool. Syst. Evol. Res.* **49**, 89 (2011).
- [3] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press (1998); J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, Massachusetts (2005).
- [4] L.L. Cavalli-Sforza, *Genes, Peoples et Langues*, Odile Jacob, Paris (1996).
- [5] L.L. Cavalli-Sforza and W.F. Bodmer, *The Genetics of Human Populations*, San Francisco: W.H. Freeman and Co (1977); W.M. Brown et al., *J. Mol. Evol.* **18**, 225 (1982); R.L. Cann, M. Stoneking and A.C. Wilson, *Nature* **325**, 31 (1987); S. Pääbo, J.A. Gifford and A.C. Wilson, *Nuc. Acid. Res.* **16**, 9775 (1988); L.L. Cavalli-Sforza, P. Manozzi and A. Piazza, *Science* **259**, 639 (1993); A.M. Macedo and S.D.J. Pena, *Parasitology Today* **14**, 119 (1998); F.M. Salzano et al, *Am. J. Hum. Biology* **11**, 359 (1999); F.M. Salzano, *Proc. Nat. Acad. Sci. (USA)* **97**, 5317 (2000).
- [6] M. Swadesh, *Int. J. Am. Linguistics* **16**, 157 (1950); — **21**, 121 (1955); V. Levenshtein, *Cybernetics and Control Theory* **10**, 707 (1966); T.V. Gamkrelidze and V.V. Ivanov, *Indo-European migration to Asia Minor*, Studies in General and Oriental Linguistics, Tokyo (1970); —, *Trends in Linguistics, Studies and Monographs* **80**, Mouton de Gruyter (1995); S. Wichmann and A.P. Grant (eds.), *Approaches to Linguistic Diversity Commemorating the Centenary of Birth of Morris Swadesh*, *Diachronica* **27.2** (special issue) (2010).
- [7] V. Schwämmle and P.M.C. de Oliveira, *Physica* **A388**, 2874 (2009); P.M.C. de Oliveira, A.O. Sousa and S.Wichmann, *On the disintegration of (proto)languages*, *International Journal for the Sociology of Language*, special issue on Language Competition and Linguistics Diffusion: Interdisciplinary Models and Case Studies (scheduled for 2013).
- [8] E.W. Holman, S. Wichmann, C.H. Brown, V. Velupillai, A. Müller and D. Bakker, *Folia Linguistica* **42**, 331 (2008); S. Wichmann, E.W. Holman, D. Bakker and C.H. Brown, *Physica* **A389**, 3632 (2010).
- [9] B. Derrida, A.J. Bray and C. Godrèche, *J. Phys.* **A27**, L354 (1994); B. Derrida, V. Hakim and V. Pasquier, *Phys. Rev. Lett.* **75**, 751 (1995); B. Derrida, P.M.C. de Oliveira and D. Stauffer, *Physica* **A224**, 604 (1996).
- [10] N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).